Math 360: Supplementary Thoughts

2013-12-03

Andrew Ross

Math Department, Eastern Michigan University

Contents

[Preliminaries 2](#_Toc365986155)

[Proposal/Project Guidance 2](#_Toc365986156)

[Project Ideas 7](#_Toc365986157)

[Related books 9](#_Toc365986158)

[Chapter 1 10](#_Toc365986159)

[Chapter 2 11](#_Toc365986160)

[Chapter 2.3: Comparative Experiments 12](#_Toc365986161)

[Chapter 3 14](#_Toc365986162)

[Chapter 4 15](#_Toc365986163)

[Chapter 5 17](#_Toc365986164)

[Chapter 6 17](#_Toc365986165)

[Chapter 7 17](#_Toc365986166)

[Chapter 7.5: Binomial and Geometric 19](#_Toc365986167)

[Chapter 8: Sampling Distributions 23](#_Toc365986168)

[Chapter 9 28](#_Toc365986169)

[Chapter 10 28](#_Toc365986170)

[Chapter 11 31](#_Toc365986171)

[Chapter 12 32](#_Toc365986172)

[2-sample z-test for proportions but paired (dependent) rather than independent 32](#_Toc365986173)

[Chapter 13 32](#_Toc365986174)

[Calculus Supplement 33](#_Toc365986175)

[Computer Science Majors 34](#_Toc365986176)

[Review 35](#_Toc365986177)

[Chapter XKCD 36](#_Toc365986178)

# Preliminaries

I always want to hear your thoughts on the class so far, so I created an online form for anonymous feedback. Please use it to let me know what you think (or if you are not concerned about anonymity, just send an email) :  
<https://docs.gooREMOVETHISgle.com/spreadsheet/embeddedform?formkey=dDF2U2djSDAtRmFUM0NsTllmOFJ2ZFE6MQ>  
You can use this throughout the semester.

Someone asked me if the random number generator in Excel is seedable (that is, can it be set to start at the same sequence over and over). There's no interface for doing that, but I researched the algorithm that the random number generator uses, and I've implemented it in simple formulas in a posted spreadsheet. You may ignore this if you want.

# Proposal/Project Guidance

For project proposals, a few things to remember:  
\* You may do a solo project or be in a team of 2 (arranged with whomever you want), but no teams bigger than 2.

\* if anyone is looking for a partner, please let me know and I will do my best to play eHarmony. But, it's not mandatory to work in a team if you don't want to.  
\* There is no competition for project topics. Multiple people or teams may do the same project.  
\* I STRONGLY ENCOURAGE you to chat with me about project ideas well before the proposal deadline, either in person or by email.  
\* After the chatting is done, feel free to email me a draft of a proposal (and perhaps a data set) for some informal feedback.  
\* See below some sample project titles from last year's Math 360 class.  
  
Proposals will generally be 1 to 2 pages, and will contain:  
\* title of project  
\* author(s) names  
\* a description of the problem you are facing  
\* a description of the available data or data collection plans  
\* a description of the proposed analysis  
\* for many projects, no literature search is needed. Others may benefit from a literature search, and may get bonus points for doing one. Giving proper credit to your sources of information or ideas is always required.  
\* If you already have the data, include a spreadsheet of it as a 2nd file when you upload the presentation  
  
If your project idea depends on getting a data set from your boss at work, you need to have the data set in hand by the time of the proposal. I've had a few projects go bad when a boss doesn't come through with promised data.  
  
A proposal does not lock you in to a topic or analysis method. If your project is not working out, contact me immediately and together we will find a new project topic.

A general tip as you're doing your projects: doing confidence intervals is almost always better than doing a hypothesis test, because a CI can be converted into an HT very easily in your head (did the CI include zero? for example), but knowing the results of an HT doesn't give you much info about the related CI. There are some projects where a CI isn't applicable, though--often those related to Chapter 12 (chi-squared tests).  The one nice thing about HT, though, is that you get a P-value, while the CI just lets you know that it was < 0.05 or whatever value you used for a CI.

It's important to create artificial data sets similar to what your real data set is. That way, when you do your processing, you can tell if you are getting what you expect to get--it's a way of debugging. You start by copying the file with your original data set, then in that copy, replacing the original data with artificial data. Then you do your analysis on the artificial data. Once you've done it for artificial data, you should be able to save *another* copy of that file, then paste your real data in where the artificial data is, and have all the calculations automatically update. This is vastly better than trying to re-create the formulas in a new sheet, since that could introduce new bugs.  
To generate a Standard Normal in excel, use  
=norminv(rand(),0,1)  
To generate a non-standard normal with mean 5 and std.dev. 3, use  
=norminv(rand(),0,1)\*3 + 5  
Another big advantage of creating artificial data is then you can compute how much your output measurements change just due to random chance, by running a whole bunch of random trials.

Your final report should be a roughly 5-to-10-page technical report (a Word file, usually). I don’t count pages, though, so don’t worry about the exact length. Please use the HomeHealthCare.doc file that I will email out as a template (remove their content, type in your own content).

Please upload both your report file and your Excel file at the same time. But, your report should have copies of any relevant figures; don't just say "see the Excel file".  
  
If you are part of a team project, \_each\_ person should upload a copy of the presentation and report.

For your final presentation, you have 2 options:

* A 5-minute Powerpoint-style presentation that you stand up and give to the class (roughly 5 slides), or
* A poster presentation, which often consists of about 12 Powerpoint slides, printed out on paper and taped to the wall of the classroom (don’t buy/use posterboard).

Each person or team of 2 may decide whether they want to do a poster or oral presentation. Either way, presentation materials should be uploaded to a dropbox inside EMU-online.

Please do not feel obligated to dress up for our presentation day in Math 360. Anyone who does dress up will be a few standard deviations from the mean, as statisticians say. Either way, it will not affect your grade at all.

However, it is important to present in a professional way (aside from how you are dressed). If I write a letter of recommendation for you, I want to be able to say how polished your presentation was—not just your slides, but your manner of speaking. This can be especially important for future teachers. In a letter of recommendation I would hope I could say “While I’ve never observed \_\_\_\_ as they teach an actual class, their final project presentation in Math 360 convinces me that they have the presentation skills to be a great teacher.”

I will recommend that some people submit their work to the Undergraduate CLASS Project Competition (USCLAP)  
<http://homepages.dordt.edu/ntintle/usproc/USCLAP.htm>  
The writeup for that has the following page limits (all in 11-point Arial, single-spaced, 1-inch margins):

1 page for title and abstract

<=3 pages for report

1 page for bibliography, if any (optional)

<=5 pages for appendices

So you might want to format your paper that way if you’re thinking of entering the contest.

The judging criteria for that contest will be the basis of the grading system for projects:  
1.   Description of the data source (15%)  
2.   Appropriateness and correctness of data analysis (40%)  
3.   Appropriateness and correctness of conclusions and discussion (20%)  
4.   Overall clarity and presentation (15%)  
5.   Originality and interestingness of the study (10%)  
NOTE: All essential materials addressing these criteria must be in the report, not confined to the spreadsheet file.

You can see the guidelines I give to my other project-based classes (Math 319, Math 419, Math 560) at this link:  
<http://people.emich.edu/aross15/project-guides/guides.html>  
though as you can see from the above, the requirements for Math 360 are a little different because of the statistical focus.  
  
----------------------------------------  
Sample project titles from previous years:  
  
Baseball player builds and home runs  
Tennis serve accuracy  
Noll-Scully simulation of sports rankings  
Spring Training vs Regular Season  
Anchoring effect  
Finding a Piecewise Linear Breakpoint in Chemistry data  
Music participation and GPA  
double-SIDS dependencies  
Swimming times  
Salary vs. Results in NCAA Tournament  
NBA scores  
Barbie Bungee Challenge  
Incumbency advantage in elections  
Comparing Distinct Audio Points in Classical and Rock Music  
Barbie Bungee Challenge  
Airbags, seat belts, bike helmets  
Spring Training vs Regular Season  
Golden Ratio in Art  
Gender differences in SAT scores

Home health care data  
double-SIDS dependencies  
GEAR-UP survey data  
Normal distributions on Wall Street?  
NBA scores  
Naive Bayesian spam filtering  
Honors college GPAs  
Predicting Course Grades from Mid-Semester Grades  
Anchoring effect  
Salary vs. Results in NCAA Tournament  
Incumbency advantage in elections

An Analysis of Correlations between Event Scores in Gymnastics Using Linear Regression

Accounting Fraud and Benford's Law

Appointment-Based Queueing and Kingman's Approximation

Are Consumers Getting all the Coconut Chocolaty Goodness They’re Paying For?

Are regular M&Ms more variable in weight than Peanut M&Ms?

Barbie Bungee Experiment

Breaking Eggs in Minecraft

Calculus-Based Probability

Comparing the Efficiency of Introductory Sorting Algorithms

Distribution of File Sizes

Do students who score better on a test’s story problems score better on the test as a whole?

Do studying Habits affect your interest in math

Do young adults under 18 and 18 and older have the same completion rate of the 3-shot regimen for Gardasil?

Does age effect half-marathon completion time

Patterns in bulk discounts

Does having high payrolls mean you will win more Major League Baseball games?

Gardasil 3-shot vaccine completion, average number of shots

Getting Hot at the Right Time: A statistical analysis of variable relative strength in the NHL

Home Field Advantage in MLB, NFL, and NBA

How random are Michigan Club Keno and Java random numbers?

Ice Cream Sales and Temperature

Is there relationship between the length of songs at the #1 spot on the Billboard Hot 100 and their respective week at #1 in time?

Lunch vs Dinner Sales at Domino's

Math Lab demand data vs Section Enrollment by Hour

Modeling School of Choice Data in Lenawee County

Pharmacy prescription pick-up times

Piecewise-Linear Regression on Concentration / Conductivity Data

Proportion & Probability of 2-Neighborly Polytopes with m-Vertices in d-Dimensions

Ranking Types of Math Questions (Algebra-based)

Scoring Trends and Home Court Advantage in Men’s College Basketball

Skip Zone on the Sidewalk

Spaghetti Bridge and Pennies

What affects a pendulum’s behavior

# Project Ideas

Are stock prices (or percent returns) normally distributed?

GPS accuracy:  
 - by time of day  
 - by weather/day-to-day  
 - within span of a few seconds or minutes  
 - from device to device  
  
Tablet/Smartphone Accelerometer Data:  
 - accuracy at 500 Hz vs 50 Hz vs 5 Hz  
 - correlation between devices  
  
---------------------------------------------------------  
advanced work on causality:  
Judea Pearl work on causality  
Bayesian networks  
Granger causality  
---------------------------------------------------------  
Future Teacher ideas (but also open to other people)

Tennessee STAR study on small class sizes  
  
problems with estimating from pie charts  
  
parents probability of pulling kids from public schools (survey)  
  
Regression through the origin: when?  
  
Instead of a regular project, work on getting the Data Analysis electronic badge?  
  
anchoring effect

Barbie Bungee: make a bungee-cord out of rubber bands, and send a Barbie (or similar toy) plunging toward the floor. Try it with a few different lengths of cord, record how far she plunges, then forecast how many rubber bands would be needed for a 12-foot drop. You can find more info online, of course.  
  
Spaghetti Bridges: make a simple bridge of straight spaghetti (not glued into a truss), see how much weight it can hold. Repeat with wider spans and/or more strands per bridge. You can find more info online.  
  
  
---------------------------------------------------------  
mechanics of statistics:  
  
LiveRegression formulas  
  
Confidence interval on s\_e for linear regression  
  
Partial Correlation in multivariate analysis  
  
---------------------------------------------------------  
CompSci ideas:  
distribution of file sizes  
 - on a hard drive (correlated to time of creation, modification, or access?)  
 - on a web server  
 - as requested from a web server  
   
distribution of packet sizes  
 - and correlation from one to the next?  
  
distribution of time gap between packets  
 - and correlation from one to the next?  
  
spam filtering (but hard to get good data set to work on)  
  
durations of jobs on the CPU  
  
memory sizes of jobs  
  
paging policies

Network round-trip times for pings

Sleep vs Cron repetitive wakings  
  
---------------------------------------------------------  
Health care:  
  
Medicare Home Health Compare  
  
Gardasil data set:  
<http://www.amstat.org/publications/jse/v19n1/gardasil.txt>  
  
SEER cancer data set,  
<http://seer.cancer.gov/>  
(need to submit application to use it)

Sports:

Pick your favorite sport and ask a statistical question about it. Some examples:  
 \* predicting player performance from previous years (helpful first step to choosing a fantasy team)  
 \* quantifying home-field advantage  
 \* (harder) quantifying time-zone advantage  
 \* bracketology

# Related books

Excel 2010 for educational and psychological statistics : a guide to solving practical problems / Thomas Quirk.

Excel 2010 for biological and life sciences statistics : a guide to solving practical problems / by Thomas J. Quirk, Meghan Quirk, Howard Horton.  
  
Converting Data into Evidence  
A Statistics Primer for the Medical Practitioner  
DeMaris, Alfred, Selman, Steven H.

Statistics with Excel website: <http://www.real-statistics.com/>  
  
  
Little Handbook of Statistical Practice  
<http://www.jerrydallal.com/LHSP/LHSP.HTM>

<http://www.stat.berkeley.edu/~aldous/Real-World/draft_book.pdf>  
On Chance and Unpredictability: 13/20 lectures on the links between mathematical probability and the real world. David Aldous, January 2012

Statistical Reasoning in Sports  
Tabor and Franklin  
<http://bcs.whfreeman.com/sris/>  
  
<http://content.bfwpub.com/webroot_pubcontent/Content/BCS_5/SRIS/Student/Appendices/AppB_Using_Excel.pdf>

Statistics: A Guide to the Unknown  
  
Forty Studies that Changed Psychology: Exploration into the History of Psychological Research

# Chapter 1

We will use this link for the Car Insurance activity:  
<https://docs.gooREMOVETHISgle.com/spreadsheet/embeddedform?formkey=dEV4M2hZanFqRlByRVRadWJiZTFQd3c6MQ>  
  
and then later we will use this link for the Data Types activity:  
[https://docs.gooREMOVETHISgle.com/spreadsheet/embeddedform?formkey=dERFS3VncWZHbUZMNnVVcnJwcXltYlE6MA](https://docs.google.com/spreadsheet/embeddedform?formkey=dERFS3VncWZHbUZMNnVVcnJwcXltYlE6MA)  
Some additional reading is included below.  
  
To prepare for our next class, we will use the following PDF file on Random Rectangles:  
<http://www.rtmsd.org/cms/lib/PA01000204/Centricity/Domain/197/RandomRectangles.PDF>  
and you should enter your answers here before class starts:  
[https://docs.gooREMOVETHISgle.com/spreadsheet/embeddedform?formkey=dHN0SlctaDFNaTY2QWctS3RraHJZbWc6MQ](https://docs.google.com/spreadsheet/embeddedform?formkey=dHN0SlctaDFNaTY2QWctS3RraHJZbWc6MQ)  
  
----------------------  
Here is some reading on the standard classifications for Data Types (nominal, ordinal, interval, ratio):  
  
<http://www.upa.pdx.edu/IOA/newsom/pa551/lecture1.htm>  
<http://en.wikipedia.org/wiki/Level_of_measurement>  
  
and an opposing viewpoint:  
[www.cs.uic.edu/~wilkinson/Publications/stevens.pdf](http://www.cs.uic.edu/~wilkinson/Publications/stevens.pdf)  
which cites, among other possible systems,  
  
Mosteller and Tukey (1977 Chapter 5):  
  
Names  
Grades (ordered labels such as Freshman, Sophomore, Junior, Senior)  
Ranks (starting from 1, which may represent either the largest or smallest)  
Counted fractions (bounded by zero and one. These include percentages, for example.)  
Counts (non-negative integers)  
Amounts (non-negative real numbers)  
Balances (unbounded, positive or negative values).  
  
Also see  
<http://stats.stackexchange.com/questions/539/does-it-ever-make-sense-to-treat-categorical-data-as-continuous>

Remember that dotplots can tell us:  
\* where the data is centered  
\*how widely spread the data is  
\* the shape of the distribution: (concentrated at an endpoint? Or in the middle?  
\* Any unusual features: gaps or outliers  
so if you're writing sentences describing a dotplot, you should write at least one sentence for each of those bullet points.  
  
For example, for the sibling-count dotplot:  
The center of the data is around 3.  
The data is spread from 1 to 9.  
The data is concentrated near the low end.  
There are no unexpected gaps or outliers.

# Chapter 2

Confounding vs Lurking:

[http://www.virmanimath.com/start-page-2012-2013/ap-stats-2012-2013/chapter-2/apstatonlineclass/confounding-and-lurking-variables](http://www.google.com/url?q=http://www.virmanimath.com/start-page-2012-2013/ap-stats-2012-2013/chapter-2/apstatonlineclass/confounding-and-lurking-variables&usd=2&usg=ALhdy28TyXMDDRSfxUuVX85clSqpyEhvfg)

Quick activity: name that sampling method  
a. Roll a die to pick a row in class, then ask each student in that row; do it twice  
b. Pick a student and then every 5th student after that  
c. Ask one student from each row  
d. pick some students, say "you guys look like typical students"  
e. throw an object, see who it hits  
f. Number the students, etc.

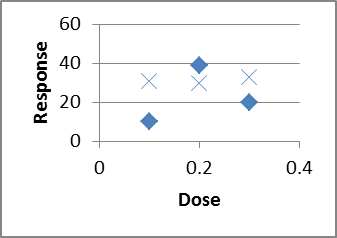
In \_\_\_\_\_\_\_ sampling, ALL groups (strata? clusters?) are used, and SOME individuals in each are sampled.  
In \_\_\_\_\_\_\_ sampling, SOME groups (strata? clusters?) are used, and ALL individuals in each are sampled.  
  
Dotplots of Random Rectangles results, look for bias  
  
Does the needed sample size grow as the population grows?  
Page 43 says no!!!!!!!

Here is a blank copy of Table 2.1 ; try to fill it out with “yes” and “no” entries by reasoning about each situation.

|  |  |  |
| --- | --- | --- |
| Study Description | Reasonable to generalize conclusions about group to population? | Reasonable to draw cause-and-effect conclusion? |
| Observational study with sample selected at random from population of interest |  |  |
| Observational study based on convenience or voluntary response sample |  |  |
| Experiment with groups formed by random assignment of individuals or objects to experimental conditions | (no entry; this row is just a header for the next 2 rows) | (no entry) |
| \* Individuals or objects used in study are volunteers or not randomly selected |  |  |
| \* individuals or objects are randomly selected |  |  |
| Experiment with groups not formed by random assignment to experimental conditions |  |  |

Chapter 2.3: Comparative Experiments  
  
"explanatory" variables are sometimes called "independent", and

"response" variables are often called "dependent",  
but in later chapters we will learn this can cause confusion.  
  
Blocking: means "putting into groups or blocks", rather than "obstructing".  
  
Blocking activity: email from a friend in the Health school here at EMU:  
> Dr. Ross,  
> I hope your holiday went well. I have recently completed a project  
> identifying some basic variables to be used in preliminary  
> evaluation of gait interventions to determine whether the new  
> intervention would be worth conducting an in-depth study about. The  
> variables include stride length, step width, stride variability, and  
> lateral displacement of the total body center of mass. As you can  
> see, these variables represent some of the most basic aspects of  
> stability, which is what we are always trying to improve or  
> maintain, and efficiency.  
>  
> My question to you is: what sample size e.g. 10 trials, 20 trials, 50  
> trials, would we need to take from both a control group (barefoot or  
> with regular shoes) and the experimental group (the intervention) in  
> order to obtain a confidence level to say that very small changes  
> between the two groups is statistically significant? An example  
> would be that the average lateral sway during gait was 5 mm less in  
> the experimental group form the control group, is that significant  
> or not?  
  
  
  
Replication  
Why do we try to do more than one trial at each level of the explanatory variable?  
Imagine this data set:



What if we had only done one trial at each dose?

Might see just the diamonds, or just the Xs.

And that's just by doing two rather than one at each level!  
Replication allows us to quantify the variability/uncertainty at each level.  
Also, when designing, choose 3 or more X values, so we can detect nonlinearity.

Controls: Positive and Negative  
Bio/Chem: when trying to detect a chemical in a sample (pollution in a lake?),  
run your procedures on some known pure water (Negative Control),  
and on some water with a known amount of pollutant deliberately added to it (Positive Control).

Computer Science: when testing spam-filtering software,  
run it on some known non-spam ("ham")--negative control,  
and on some known spam -- positive control.  
  
What is the difference between placebo and control?  
Placebos are meant to fool \_people\_--usually unnecessary on non-people.  
While control experiments apply to people and non-people alike.  
But you should still handle animals in the control group the same way (incl. surgery?)

Bias in cancer screening:

<http://www.cancer.gov/ncicancerbulletin/112712/page4>  
  
Crunching Numbers: What Cancer Screening Statistics Really Tell Us, by [Sharon Reynolds](http://www.cancer.gov/ncicancerbulletin/bios/reynolds)

# Chapter 3

A very interesting data set/set of histograms, which we would expect to have a Normal distribution like the SAT or ACT, but it's definitely non-normal in very interesting ways: (you'll need to scroll down past the description of how he collected the data in a sneaky way):  
<http://deedy.quora.com/Hacking-into-the-Indian-Education-System>  
whereas SAT data can be found at  
<http://research.collegeboard.org/programs/sat/data>  
<http://research.collegeboard.org/content/sat-data-tables>

# Chapter 4

Here is the data that you entered this semester on your own height, in inches above 5-foot-0-inches.  
We will use this in class.  
Not to spoil the fun, but we will:  
\* make a histogram to see its general shape, in bins of width 2 inches  
(use the histogram template file we've been using)  
\* compute the mean & SD  
\* compute another histogram with bin widths of 1 SD, centered on the mean  
\* look at the % of data points within +/- 1 SD of the mean, and then +/- 2 SD of the mean, and +/- 3 SD.  
  
6  
10.5  
9  
7  
10  
8  
12  
0  
8  
10  
5  
2  
6  
10  
9.5  
4 in  
16.5  
17  
6  
9  
13  
6  
11  
9  
7.75  
6  
15  
2  
0.4  
4 inches  
6  
5  
14.5  
12  
3  
5-foot-6-inches  
4  
11  
10  
11  
4.75  
6  
12  
7  
  
8  
9  
  
12.5  
  
8  
  
  
And here's another data set, on how long it took students in one of my Math 110 classes to walk from Green Lot to Pray-Harrold, in decimal minutes:  
9.4166666667  
8.35  
9.6  
10.4666666667  
10.5333333333  
12.15  
7.7666666667  
8.1666666667  
6.3666666667  
6.3666666667  
0.1334166667  
8.2166666667  
7.25  
7.3333333333  
22.7833333333  
7.55  
6.5833333333  
4.45  
7.1  
3.7166666667  
5.7166666667  
11.9833333333  
6.2666666667  
8.6333333333  
5.7  
7.25  
7.9666666667  
7.8833333333  
6.766666666

# Chapter 5

<http://en.wikipedia.org/wiki/Regression_to_the_mean>  
<http://en.wikipedia.org/wiki/Ecological_fallacy>

Why best cannot last: Cultural differences in predicting regression toward the mean  
<http://onlinelibrary.wiley.com/doi/10.1111/j.1467-839X.2010.01310.x/abstract>  
Roy R. Spina1, Li-Jun Ji2,\*, Michael Ross3, Ye Li, Zhiyong Zhang5  
Article first published online: 16 AUG 2010  
  
DOI: 10.1111/j.1467-839X.2010.01310.x  
  
Keywords:  
culture; lay theories of change; prediction; regression toward the mean  
Four studies were conducted to investigate cultural differences in predicting and understanding regression toward the mean. We demonstrated, with tasks in such domains as athletic competition, health and weather, that Chinese are more likely than Canadians to make predictions that are consistent with regression toward the mean. In addition, Chinese are more likely than Canadians to choose a regression-consistent explanation to account for regression toward the mean. The findings are consistent with cultural differences in lay theories about how people, objects and events develop over time.

# Chapter 6

How can you analyze this study?  
<http://www.healthline.com/health-news/children-autism-risk-appears-early-in-the-placenta-042513>  
  
Autism Risk Detected at Birth in Abnormal Placentas  
Written by Julia Haskins | Published April 25, 2013

# Chapter 7

We might also have a quick quiz in class about how shifting or scaling affects the mean, variance, SD, IQR, etc., and the proper formula for the sample variance.  
  
In class, we used the dotplot-histogram-crf-1000 sheet to investigate questions like:  
\* Is E[X+Y] = E[X] + E[Y] ?  (yes, it always is--doesn't even need independence!)  
\* Is Std(X+Y) = Std(X) + Std(Y) ? (no, it basically never is!)  
\* Is Var(X+Y) = Var(X) + Var(Y) ? (in Excel it was close enough; with infinite trials, it's exactly true,  
but we need to require that X and Y be independent, or at least uncorrelated)  
  
Some other questions we could ask:  
\* Is E[X^2] = ( E[X] )^2 ?  
\* Is E[1/X] = 1/ E[X] ?  
\* Is E[X\*Y] = E[X]\*E[Y] ?  
  
If you look at the 2nd multi-plotting sheet inside that file I sent, you will see a copy of what we did today, and some experiments as suggested above.  
  
In other news, here is some advice on keeping notation straight:  
  
<http://heather.cs.ucdavis.edu/~matloff/132/PLN/ProbStatBook.pdf>  
Page 15:  
Do not write/think nonsense. For example: the expression "P(A) or P(B)" is nonsense--do you see why? Probabilities are numbers, not boolean expressions, so "P(A) or P(B)" is like saying, "0.2 or 0.5" -- meaningless.  
  
Similarly, say we have a random variable X. The "probability" P(X) is invalid. P(X = 3) is valid, but P(X) is meaningless.  
  
Please note that = is not like a comma, or equivalent to the English word therefore. It needs a left side and a right side; "a = b" makes sense, but "= b" doesn't.  
  
Similarly, don't use "formulas" that you didn't learn and that are in fact false. For example, in an expression involving a random variable X, one can NOT replace X by its mean. (How would you like it if your professor were to lose your exam, and then tell you, "Well, I'll just assign you a score that is equal to the class mean"?)  
  
And, from Rossman and Chance, "Brief Review of Set Operations and Properties":  
An event is a set, while a probability is a number.  
One calculates probabilities of events (and therefore of sets), but probabilities are numbers. The following \_meaningless\_ statements are examples of nonsensical confusions of sets and numbers:  
P(A) intersect P(B)  
(P(A))'  
P(1-A)  
P(A+B)  
  
Examples of \_meaningful\_ statements about events and probabilities include:  
P(A intersect B)  
P(A')  
1-P(A)P(B)+P(B)

Chapter 7.5: Binomial and Geometric  
  
Classic question: if you flip a (unfair?) coin n times, how many times will it come up heads?  
The # of heads has a Binomial distribution.  
2nd classic questions: if you flip a (unfair?) coin \_until\_ you get your first heads,  
how many flips will it take?  
The # of flips has a Geometric distribution.  
  
Binomial has a fixed # flips, random #heads  
Geometric has a random #flips, fixed #heads (just 1)  
  
We already saw: P(1 energy-efficient fridges out of 3)  
involved 3 different outcomes (3 trials, choose 1 E fridge)  
  
Binomial PMF: one thing Prof. Casey has identified as  
"something to know cold"!  
Book uses p(x) but then also uses p for success probability--dangerous!  
nCx \* p^x \* (1-p)^(n-x); x=0,1,..., n  
Binomial PMF applet:  
<http://www.distributome.org/js/calc/BinomialCalculator.html>  
  
Example problems:  
n=10 fair coin flips (p=1/2), P(X=5)?  
0.24609375  
Can use BINOMDIST(x,n,p, false) for PMF in Excel  
What about P(X<=3 )?  
Could do P(X=0)+P(X=1)+P(X=2)+P(X=3)  
but there's a better way: binomdist using the cumulative=true option:  
BINOMDIST(x,n,p,true) = P(X<=x )  
0.171875  
What about P(X>7)? There's no reverse-cumulative option.  
Instead, say: P(X>7) is the opposite of P(X<=7: P(X>7)=1-P(X<=7)  
then do 1-binomdist(7,n,p,true)  
What about P(X>=7) ? Change to X>6, then use 1-P(X<=6)  
  
How about P(3 < X <= 8)? Change to P(X<=8) - P(X<=3)  
  
Mean & Standard Deviation:  
If you flip a 60% coin 10 times, how many H do you expect? 6, of course.  
So mean # of successes is E[X]=n\*p  
StdDev isn't so obvious. Var(X)=n\*p\*(1-p)  
This is much more useful in Chapter 7.8, Binomial Approximations.  
  
Geometric:  
# of FLIPS until success:  
P(X=x) = failure on x-1 flips, success on 1 flip = (1-p)^(x-1) \* p  
x=1,2,3,...  
Some books call this the G1 distribution since it starts at x=1.  
If we asked # FAILURES, not #FLIPS, that would start at x=0, call it G0.  
(PMF is slightly different)  
Wikipedia shows both types:  
<http://en.wikipedia.org/wiki/Geometric_distribution>  
P(X<=x)=1-P(X>x)=1-P(x failures at start)=1-(1-p)^x  
No equivalent function in Excel.  
  
Geometric distribution applet:  
<http://www.distributome.org/js/calc/GeometricCalculator.html>  
  
Book skips: Mean & Var for Geom  
If each coin flip has a 1-in-10 chance of H, how many flips until H?  
10 is the obvious answer, and it's right: E[X] = 1/p  
Var(x)=(1-p)/p^2   
  
Big important property of Geometric distribution: Memoryless!  
If E[X]=10 and we're already on flip #8 without a H, E[#remaining flips]=10  
not 10-8=2  
  
Chapter 7.6: Normal Distribution  
If you want to see the formula for the bell curve, visit  
<http://en.wikipedia.org/wiki/File:10_DM_Serie4_Vorderseite.jpg>  
We hardly ever use that formula in Stats class, though, other than  
to graph it and shade in some areas so we can see what we are doing.  
  
Start with Standard Normal: mean=0, stddev=1  
This is so special it gets its own letter: z instead of x  
(we already calculated z-scores; it's not a coincidence!)  
  
<http://www.flashandmath.com/mathlets/calc/antplot/antplot.html>  
1/sqrt(2\*pi)\*e^(-x^2/2)  
y range on f(x) to [0,0.4]  
x range on f(x) to [-3,3]  
F(-3) = 0.0013  
  
It's hard to compute something like P(Z<=1) from scratch,  
so we use tables or Excel formulas.  
Shade area on bell curve for Z<=2,  
look up in table,  
use Excel formula =normdist(2,0,1,true)  
And highlight on CDF graph.  
Now try for P(-0.5 < Z < 0.5)  
  
Now backwards: what z cutoff gives P(Z<=z) = 0.80 ?  
And double-backwards: what z cutoff gives P(-z < Z < z)=0.95 ?  
  
Non-Standard distributions: translate to z-scores and back.  
On the graph printouts, write in x values next to z values.  
Speeds on a particular road average 40 mph, sigma=5  
What % of speeds are under 45?  
What % of speeds are between 30 and 50 ?  
  
Normal Distribution applet:  
<http://www.distributome.org/js/calc/NormalCalculator.html>  
(the autoscaling makes it near-worthless because it always looks the same,  
but that's kind of the point!)  
  
Excel: for finding Pr from cutoff, =normdist(cutoff, mean, std, true)  
For finding cutoff from Pr, use =norminv(prob, mean, std )  
  
Chapter 7.7: Checking for Normality, and Normalizing Transformations  
  
Big initial notice: don't sweat the details of the formulas here,  
since each book and software package does things a little differently.  
  
Normal probability Plot, also called Q-Q (Quantile-Quantile) plot for Normal  
  
Basic idea: could make a CRF plot directly from data (no binning),  
then overlay a NormalCDF plot. But it's hard to tell how well two CURVES  
match.  
So we plot x=exact Normal quantiles, y=data quantiles, which should  
make a straight line if the distribution is Normal.  
  
Show Q-Q plot in existing Excel file; don't construct by hand in class!  
  
  
If the data isn't Normal, sometimes we transform it  
(take sqrt, cubert, or log) to see if that makes it more normal.  
Some people say that stock market returns are normal once you take logs:  
ln(price today / price yesterday)  
this is called a LogNormal distribution.  
  
Read in the book: using correlation coefficient to decide if it's  
reasonably close to linear.  
  
Chapter 7.8: Approximating Binomial with Normal  
  
Important intuition:  
We noticed that the Binomial distribution often looked bell-curve-shaped.  
So we could approximate Binomial probabilities with Normal probs.  
Match the mean & the StdDev.  
mean = n\*p, stddev = sqrt(n\*p\*(1-p) )  
  
P( a < Binom(n,p) < b ) approx= P( a < Normal < b)  
  
For instance: flip 1000 times with p=0.4  
mean=1000\*0.4=400, std=sqrt(1000\*0.4\*0.6)=  
15.4919333848297  
Pr( Binom within +/- 15 of mean of 400?)  
  
Exact: =binomdist(400+15,1000,0.4,true)-binomdist(400-15,1000,0.4,true)  
Google Docs spreadsheet gives an overflow error:  
#NUM!  
name-brand Excel gives  
0.666757849  
Normal Approximation: =normdist(400+15,400,15.49,true)-normdist(400-15,400,15.49,true)  
0.667138730035761  
  
  
Less-important, detail-oriented stuff: using < versus <=,  
and the Continuity Correction.  
  
binomial approximations applet:  
<http://www.jsc.nildram.co.uk/examples/sustats/normalapproximations/NormalApproximationsApplet.html>

# Chapter 8: Sampling Distributions

We'll use the “billionaire-dotplot-histogram-crf-1000” in file in class.  
  
We'll also use these applet pages:  
<http://www.stat.auckland.ac.nz/~wild/WPRH/index.html>  
  
<http://onlinestatbook.com/2/index.html>

A very good article on why it’s important that the standard error falls inversely with sqrt(n) :

<http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>

Below, I'm including some text about sampling distributions from a different book. Please read it.  
  
  
from "Workshop Statistics, 4th Edition" by Rossman and Chance:  
Topic 13: Sampling Distributions: Proportions  
page 278  
  
Watch Out!  
  
\*The concept of sampling distribution is one of the most difficult statistical concepts to firmly grasp because of the different "levels" involved. For example, here the original observational units are the candies, and the variable is the color (a categorical variable). But at the next level, the observational units are the samples, and the variable is the proportion of orange candies in the sample (a quantitative variable). Try to keep these different levels clear in your mind.  
  
[to Math 360 students: we will see this categorical/quantitative split in Chapter 8.3; it's not so apparent in Chapter 8.1 and 8.2]  
  
\* It's essential to distinguish clearly between parameters [math 360: our book calls them population characteristics] and statistics. A parameter is a fixed numerical value describing a population. Typically, you do not know the value of a parameter in real life, but you may perform calculations assuming a particular parameter value. On the other hand, a statistic is a number describing a sample, which varies from sample to sample if you were to repeatedly take samples from the population.  
  
\* Notice that the Central Limit Theorem (CLT) specifies /three/ things about the distribution of a sample proportion [and also about a sample mean]: shape, center (as measured by the mean), and spread (as measured by the standard deviation). It's easy to focus on one of these aspects and ignore the other two. As with other normal distributions, drawing a sketch can help you to visualize the CLT.  
  
\* Ensure that these conditions hold before you apply the CLT: the sample needs to have been chosen randomly, and the sample size condition requires that n\*pi>=10 and n\*(1-pi)>=10 [math 360: we say n\*p>=10 and n\*(1-p)>=10].  ... it's the normal shape that depends on this condition. The results about the mean and standard deviation hold regardless of this condition.  
  
\* Notice that the sample size, relative to the value of the population parameter, is a key consideration when prediction whether or not the sampling distribution will be approximately normal. However, changing the sample size is in no way changing the parameters or shape of the categorical population distribution!  
  
\* As long as the population size is much larger than the sample size (say, 20 times larger), the /population/ size itself does not affect the behavior of the sampling distribuiton. This sounds counterintuitive to most people, because it means that a random sample of size 1000 from one [US] state will have the same sampling variability as a random sample of size 1000 from the entire country (with the same population proportion). But think about it: if chef Julia prepares soup in a regular-sized pot and chef Emeril prepares soup in a restaurant-sized vat, you can still learn the same amount of information about either soup from one spoonful. You don't need a larger spoonful to decide whether you like the taste of Emeril's soup.  
  
\* As we've said before, try not to confuse the sample size with the number of samples. The sample size is the important number that affects the behavior of the sampling distribution. In practice, you only get one sample. We have asked you to simulate a large number of samples only to give you a sense for how sample statistics vary under repeated sampling; we have tried to ask for enough samples (typically 500 or 1000) to give you a sense of what would happen in the long run. In fact, now that you know the Central Limit Theorem's description of how sample proportions vary under repeated sampling, you no longer need to simulate taking many samples from the population.

Simpson’s Paradox supplemental reading:

From Jeff Witmer at Oberlin:  
<http://new.oberlin.edu/dotAsset/1801848.ppt>  
  
From Tom Moore at Grinnell:  
<http://www.math.grinnell.edu/~mooret/reports/SimpsonExamples.pdf>

And an article that’s very good for pre-service teachers: Representations of Reversal:An Exploration of Simpson's Paradox by Lawrence Mark Lesser

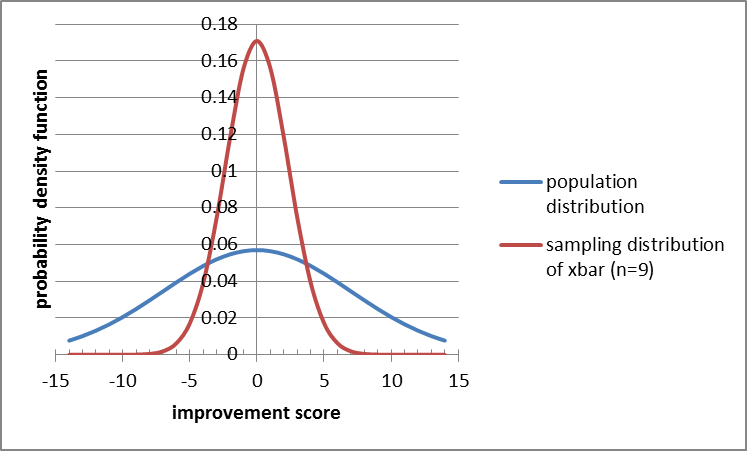
<http://www.statlit.org/PDF/2001LesserNCTM.pdf>

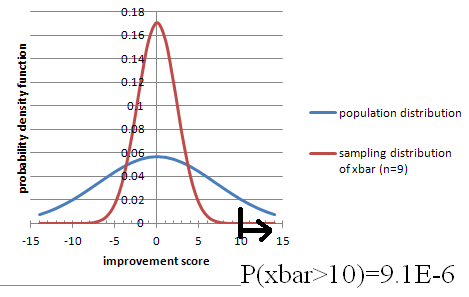
Also, this article discusses the possibility of a “Double Simpson’s Paradox”, and then it turns out that such a thing is impossible:

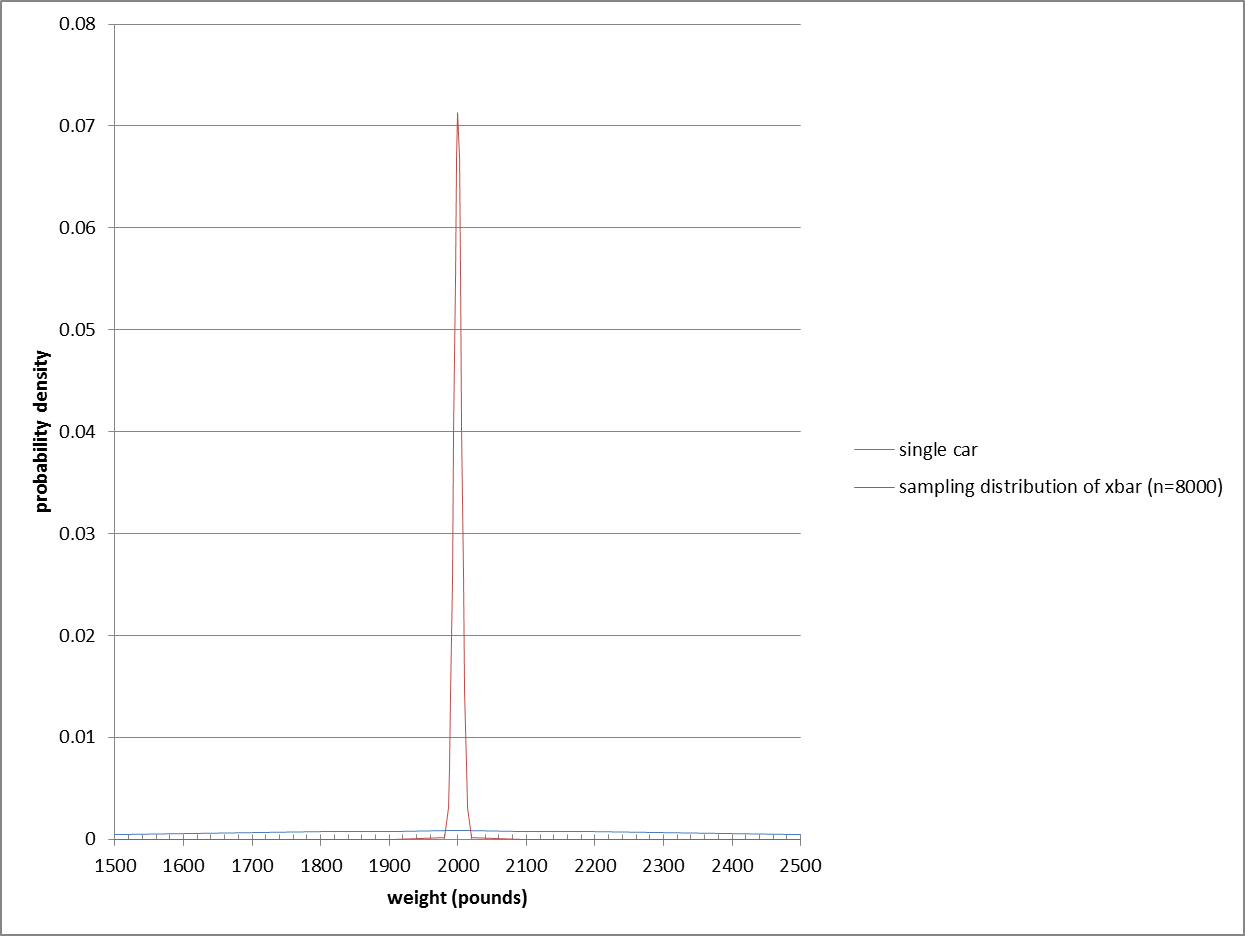
Friedlander, Richard, and Stan Wagon. "Double Simpson's Paradox." Mathematics Magazine 66 (October 1993): 268

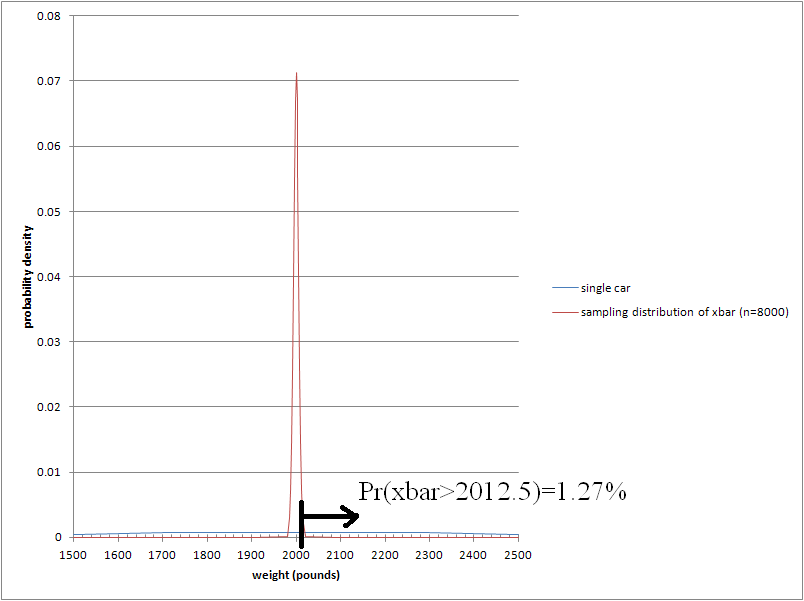
Help on Problem 8.14:

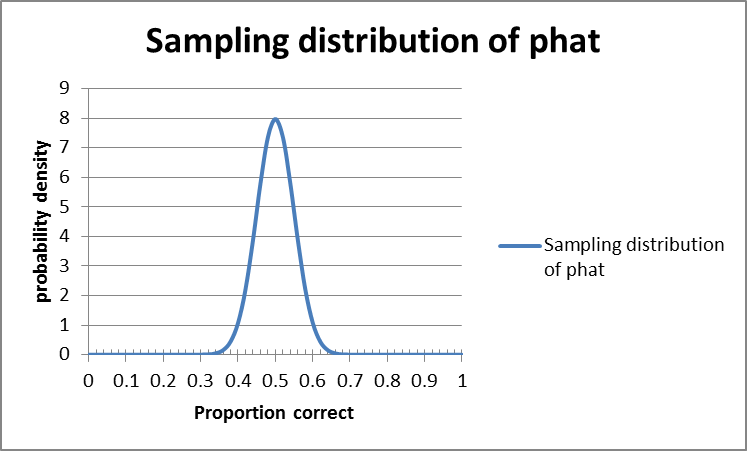
A student asked me for more guidance on homework 8 #8.14, where I ask you to draw a picture of the sampling distribution (on paper, or however you want--no need to turn it in).  
  
To give some examples of what I was imagining, I went and drew curves from our class examples (rather carefully) and they are in the file “m360-sampling-distribution-drawings.xls” . You don't have to do them this carefully at all--you could just freehand them.

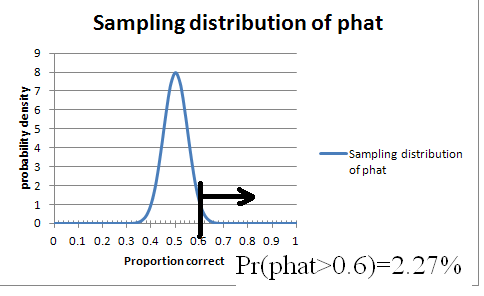
  
  
I added an illustration of which probabilities we were computing by freehanding stuff in Microsoft Paint. Again, you don't have to do that.











# Chapter 9

The applet we used last time, on sampling distribution:  
<http://onlinestatbook.com/stat_sim/sampling_dist/index.html>  
  
Algebraic proof of why we use n-1 in sample variance:  
<http://pascencio.cos.ucf.edu/classes/Methods/Proof%20that%20Sample%20Variance%20is%20Unbiased.pdf>  
<http://en.wikipedia.org/wiki/Bessel%27s_correction>  
  
Khan Academy videos and applets on use of n-1 in sample variance:  
<https://www.khanacademy.org/math/probability/descriptive-statistics/variance_std_deviation/v/simulation-showing-bias-in-sample-variance>  
<https://www.khanacademy.org/math/probability/descriptive-statistics/variance_std_deviation/v/another-simulation-giving-evidence-that--n-1--gives-us-an-unbiased-estimate-of-variance>  
  
Confidence Intervals Applet from Peck/Olsen/Devore:  
  
<http://www.thomsonedu.com/statistics/book_content/0495118737_peck/misc/SS/Applets/applets2/PeckConfsim.html>

# Chapter 10

Big disclaimer: almost everything in science uses the basic idea of this chapter, but it has fundamental problems that most statisticians acknowledge!

<http://golem.ph.utexas.edu/category/2010/09/fetishizing_pvalues.html>  
  
  
The Earth Is Round ( p<0.05)  
by Jacob Cohen  
<https://labs.psych.ucsb.edu/janusonis/skirmantas/cohen1994.pdf>  
  
<http://quomodocumque.wordpress.com/2013/05/13/tantalisingly-close-to-significance/#comment-16547>  
  
Chapter 11.9, page 230 of:  
<http://heather.cs.ucdavis.edu/~matloff/132/PLN/ProbStatBook.pdf>  
What's Wrong with Signicance Testing|and What to Do Instead  
  
book: The Cult of Statistical Signicance, by S. Ziliak and D. McCloskey  
  
quotes about HT  
<http://www.indiana.edu/~stigtsts/quotsagn.html>  
  
  
WE NEVER SAY what the probability of H0 itself is!  
That's the Frequentist way. Bayesians are happy to talk about it.  
  
In a criminal trial, we (the people) want to show strong evidence of guilt, so Ha = guilty; then H0 = not guilty, which is not the same as innocent. (though in the French system it's the opposite!)  
  
If the data is unlikely under the supposition of H0, then we "reject H0" and accept Ha.  
BUT if the data isn't particularly unlikely (supposing H0), then we NEVER NEVER NEVER "accept H0"; we just fail to reject it. In a criminal trial, we don't declare them innocent, we just fail to prove that they are guilty.  
  
  
Hypotheses are about things we DON'T know, like mu or p or sigma. There's no point having a hypothesis about something we DO know, like xbar or phat.  
  
Often we reason this way:  
I want to show my product is better than the specification:  
Ha : my quality (population % good) > specified value of p  
So what's the opposite?  
H0: my quality (population % good) <= that specified value fo p  
BUT: 1) to assume H0 is true we need a specific value, not just <=,  
AND the most conservative thing to do is let it be as close as possible to what I'm trying to show: =specified value rather than <specified.  
So while H0 often would naturally be a <= or >= we state & treat it as an =  
  
We can never show that two things ARE equal, just fail to show that they aren't.  
(but maybe we just didn't collect enough data)  
  
Which tail should you pick? Depends on what you're trying to prove:  
my product is better than the standard  
this product is worse than the standard.  
Example 10.3 is very good!  
But note that we might be willing to use the new treatment even if it is provably less effective than the old, as long as it's a lot cheaper or has less side effects.  
BUT you have to decide before you see the data; if you let your choice  
be influenced by the data, you won't hit your confidence/significance level!  
  
Wikipedia page on: Testing hypotheses suggested by the data  
<http://en.wikipedia.org/wiki/Testing_hypotheses_suggested_by_the_data>  
  
One-sided (one-tailed) vs Two-sided (two-tailed) tests:  
If Ha is "this population is better" than a standard, use > of course  
or use < if Ha is "this population is worse" than a standard.  
Sometimes we just want to show it's \_different\_ than a standard:  
Ha: p not equal to some standard value. This is called a two-sided or two-tailed test.  
Two-sided is the default if you can't decide; it's safer because its cutoff values are farther away from the hypothesized value.  
  
Type III (and IV) error:  
<http://en.wikipedia.org/wiki/Type_III_error#Type_III_errors>  
  
Page 585, Example 10.7: null hypothesis implicitly includes mu<15, but book says includes mu>15 (am I right on that?)  
  
Here is one way of laying out the 9-step process in Excel, all in one row (or actually two, one for headings and the other for numbers/calculations):  
  
definition        H0 value of p        Ha        alpha        pop. Size (approx)        n        #successes        phat        n\*p        n\*(1-p)        sample/pop        SE of phat ASSUMING H0        test statistic z        p value        decision

Reading Prompts for a Concept-based quiz on CI and HT:

Very little calculation is required for this quiz; it's more about phrasing.  
  
Here are the situations that will be part of the quiz. Each situation will be followed by 5 options, and you will chose 1 of those 5.  
  
1. Only 33% of students correctly answered a difficult multiple-choice question on an exam given nationwide. Professor Chang gave the same question to her 35 students, hypothesizing that they would do better than students nationwide. Despite the lack of randomization, she performed a one-sided test of the significance of a sample proportion and got a P-value of 0.03. Which is the best interpretation of this P-value?  
  
  
 2. Researchers constructed a 95% confidence interval for the proportion of people who prefer apples to oranges. They computed a margin of error of +- 4%. In checking their work, they discovered that the sample size used in their computation was 1/4 of the actual number of people surveyed. Which is closest to the correct margin of error?  
            
  
 3. A survey of 200 randomly selected students at a large university found that 105 favor a stricter policy for keeping cars off campus. Is this convincing evidence that more than half of all students favor a stricter policy for keeping cars off campus?  
            
  
 4. In college populations, the annual incidence of infectious mononucleosis has been estimated to be as many as about 50 cases per 1000 students. A university student health service took a survey of students to test whether the rate of mononucleosis on their campus is different from this national rate. With alpha=0.05, they rejected the null hypothesis. Which is the best interpretation of "alpha=0.05" in this context?  
Source: [www.aafp.org/afp/20041001/1279.html](http://www.aafp.org/afp/20041001/1279.html)  
            
  
 5. In a pre-election poll, 51% of a random sample of voters plan to vote for the incumbent. A 95% confidence interval was computed for the proportion of all voters who plan to vote for the incumbent. What is the best meaning of "95% confidence"?  
            
  
 6. Sheldon takes a random sample of 50 U.S. housing units and finds that 30 are owner occupied. Using a significance test for a proportion, he is not able to reject the null hypothesis that exactly half of U.S. housing units are owner occupied. Later, Sheldon learns that the U.S. Census for the same year found that 66.2% of housing units are owner occupied. Select the best description of the type of error in this situation.

# Chapter 11

Here's a video that is meant for a different textbook, but it's still a good overview of what we've been looking at recently. It's interesting to note that their requirements for tests are slightly different: n>40 instead of n>30 for non-normal data, and n\*p>15 in some cases, >10 in others, and >5 in some!  
  
<http://bcs.whfreeman.com/statsportal/bps4e/stattutor/chapter%2022/Inferences%20about%20Variables%20Part%20III%20Review.swf>

# Chapter 12

In class:  
Let's see if we can gather some binomial data. Let's use families with 2 kids (if your family has more than 2, just consider the first 2).  
Go to  
<https://docs.gooREMOVETHISgle.com/forms/d/1NBh5OpRoXPXQMy1VcpjTIuRKeQ5tryEFI3sIMnAZ5b0/viewform>  
and enter either GG, GB, BG, or BB.  
---------------------------------  
Chi-squared stuff:  
  
Desmos calculator that I made:  
<https://www.desmos.com/calculator/swmkwk8fyx>  
  
Distributome:  
<http://www.distributome.org/js/calc/ChiSquareCalculator.html>  
  
My excel sheet with a slider; might work only on PCs rather than Macs. Also, requires you to enable macros, which is dangerous in general but probably safe for this one file.  
-------------------------------------  
  
I'm also attaching two data sets on population by city or county, that we might have time to analyze in class using Benford's law:  
Log10(1+1/i) for first digit i.

## 2-sample z-test for proportions but paired (dependent) rather than independent

<http://www.uvm.edu/~dhowell/methods8/Supplements/Testing%20Dependent%20Proportions.pdf>  
  
<https://onlinecourses.science.psu.edu/stat504/node/96>  
  
<http://home.ubalt.edu/ntsbarsh/business-stat/otherapplets/PairedProp.htm>

# Chapter 13

<http://analyse-it.com/docs/220/Resources/Images/ScatterPlot.PNG>

Testing for linearity:

<http://en.wikipedia.org/wiki/Lack-of-fit_sum_of_squares>

# Calculus Supplement

See that separate file.

Leemis diagram of distributions:

<http://www.math.wm.edu/~leemis/chart/UDR/UDR.html>

Decision tree of what distribution to use:

<http://pages.stern.nyu.edu/~adamodar/New_Home_Page/StatFile/statdistns.htm>

College Math Journal, Vol 31 No 4 September 2000: The Lognormal Distribution  
by Brian E Smith and Francis J Merceret  
page 259-261

<http://classes.soe.ucsc.edu/ams113/Winter03/>  
classes.soe.ucsc.edu/ams113/Winter03/calc.pdf  
  
<http://www.zweigmedia.com/RealWorld/cprob/cprobintro.html>  
  
http://cs.gettysburg.edu/~leinbach/DRVTI\_Conf/Kellett.PDF  
  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2931677/>  
  
  
<http://jwilson.coe.uga.edu/EMAT6680Fa06/Sexton/STAT%206070/Web2.html>  
  
Activites for Calc-Based Statistics Classes  
<http://www.mathspace.com/NSF_ProbStat/Teaching_Materials/Primarily_Statistics.htm>

# Computer Science Majors

I asked some professors in our CompSci department what CS majors should get out of Math 360, and here are some of their responses:

FDR: False Discovery Rate  
<http://en.wikipedia.org/wiki/False_discovery_rate>  
  
p-values and q-values:  
<http://www.nonlinear.com/support/progenesis/samespots/faq/pq-values.aspx>  
-------------------------------------  
I think it would be helpful if CS students knew about density/distribution functions, perhaps with more emphasis on discrete (but not entirely).  
Conditional probability , Bayes rule , joint densities.  
  
Possible Projects:  
Diagnosis: medical, mechanical, really anything. Use Naive Bayesian Inference  
Classification: Naive Bayesian Inference with MAP estimator - spam/textual filter classifier  
Simulation: Monte Carlo techniques, various types of Markov processes  
Hypothesis Testing - did this type of user interface increase productivity, did that new protocol increase through-put, etc.  
----------------------------  
In some cases, it is a bit hard to deconvolute the stats from the science.  
  
Recommender systems has gone the way of matrix decompositions, but understanding distributions is definitely important.  Variance, standard deviation, covariance... different forms of correlation, mutual information. Also, ways of looking at error of prediction  
  
Bioinformatics is a bit different, since it is a bit more experimental. you see a lot of use of Fisher's exact test for testing "enrichment" of annotations --- e.g., does a set of n genes found experimentally include more members who are annotated to appear in the nucleus than you would expect by chance? Understanding p-values, q-values/FDR is important.  
  
Bayes rule shows up in several settings.  
  
  
-----------------------------  
Personally, I would think that some Bayesian inference could be useful.  Maybe signal detection theory.  Cluster analysis might fit MATH 360.  You would find a lot of potential projects in artificial intelligence, pattern recognition, and machine learning, among other areas.  
  
Signal detection theory:  
<http://psych.hanover.edu/JavaTest/STD/>  
<http://wise.cgu.edu/sdtmod/>  
<http://en.wikipedia.org/wiki/Detection_theory>  
<http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html>

# Review

Try to make a summary table with these column headings:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Row# | #samples : 1, 2, >=3 | means or proportions | paired or not | # tails | CI or HT? | Excel function |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |

The following web site shows you a bunch of statistics scenarios and you click on the statistical technique that is most appropriate, and get instant feedback. There's one type of test we haven't talked about, though: ANOVA. When clicking on types of tests to include, don't click on ANOVA.  
  
<http://www-personal.umich.edu/~hamms/spotlight/nts2.html>  
  
PS: if you're curious about ANOVA:  
" In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups."  
<http://en.wikipedia.org/wiki/Analysis_of_variance>

See the scanned pages with a lot of good questions, most of which are concept-based. Some notes:  
\* Only work on the ones that are multiple-choice; the ones labeled "Investigation" are not part of the practice test.  
\* There are occasional problems that require you to do some statistical calculations, including using z, t, or chi-squared tables or Excel functions.  
\* on page 621 of the scanned pages, you may skip problem C6 (a six-sided die)  
\* on page 663 of the scanned pages, you may skip problem C1.  
  
The concept test that we did via emu-online, on confidence intervals and hypothesis tests, is also very good for you to study from.

However, the actual test questions will not be simple alterations of these questions; they will be new contexts.  
  
The test will also include a few computational questions; here are some practice problems for those:  
  
A poll of 1000 people found that 53% said they were Republican and 47% said Democrat (we are ignoring unaffiliated voters). Of the Republicans, 20% were in favor of a particular new political proposal. Of the Democrats, 18% were in favor. Do an appropriate statistical analysis; show all work and reasoning.  
  
  
In planning for a wind turbine to generate electric power, a city put up a wind-speed sensor in Location A and collected 7 days of data, with resulting speeds (avg per day, in mph) of:  
10 12 11 12 9 9 12  
The sensor was then moved to Location B, whose measurements were then  
9 11 13 12 12 9 8  
Do an appropriate statistical analysis; show all work and reasoning. If you need to make any assumptions, write down what you are assuming.

# Chapter XKCD

Null Hypothesis:  
<http://xkcd.com/892/>  
  
Multiple Testing:  
<http://xkcd.com/882/>  
  
Frequentist vs Bayesian:  
<http://xkcd.com/1132/>  
  
Statistically Significant outlier  
<http://www.xkcd.org/539/>  
  
Cell Phones and Cancer:  
<http://www.xkcd.org/925/>